GASP: A pan-specific predictor of family 1 glycosyltransferase specificity enabled by a pipeline for substrate feature generation and large-scale experimental screening

David Harding-Larsen[1,§], Christian Degnbol Madsen[1,2,§], David Teze[1], Tiia Kittilä[1], Mads Rosander Langhorn[1], Hani Gharabli[1], Mandy Hobusch[1], Felipe Mejia Otalvaro[1], Onur Kırtel[1], Gonzalo Nahuel Bidart[1], Stanislav Mazurenko[3,4], Evelyn Travnik[1], Ditte Hededam Welner[1,*]

[1]The Novo Nordisk Center for Biosustainability, Technical University of Denmark, Kemitorvet 220, Kgs. Lyngby DK-2800, Denmark
[2]Melbourne Integrative Genomics, Schools of BioSciences and of Mathematics & Statistics, University of Melbourne, Building 184 Parkville, VIC, 3010, Australia
[3]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Karmenice 5, 625 00 Brno, Czech Republic
[4]International Clinical Research Center, St. Anne's University Hospital Brno, Pekarska 53, 656 91 Brno, Czech Republic

[§]Contributed equally to this work
*Corresponding author: diwel@biosustain.dtu.dk

Abstract

Glycosylation represents a major chemical challenge; while it is one of the most common reactions in Nature, conventional chemistry struggles with stereochemistry, regioselectivity and solubility issues. In contrast, family 1 glycosyltransferase (GT1) enzymes can glycosylate virtually any given nucleophilic group with perfect control over stereochemistry and regioselectivity. However, the appropriate catalyst for a given reaction needs to be identified among the tens of thousands of available sequences. Here, we present the Glycosyltransferase Acceptor Specificity Predictor (GASP) model, a data-driven approach to the identification of reactive GT1:acceptor pairs. We trained a random forest-based acceptor predictor on literature data and validated it on independent in-house generated data on 1001 GT1:acceptor pairs, obtaining an AUROC of 0.79 and a balanced accuracy of 72%. GASP is capable of parsing all known GT1 sequences, as well as all chemicals, the latter through a pipeline for the generation of 153 chemical features for a given molecule taking the CID or SMILES as input (freely available at https://github.com/degnbol/GASP). GASP had an 83% hit rate in a comparative case study for the glycosylation of the anti-helminth drug niclosamide, significantly outperforming a hit rate of 53% from a random selection assay. However, it was unable to compete with a hit rate of 83% for the glycosylation of the plant defensive compound DIBOA using expert-selected enzymes, with GASP achieving a hit rate of 50%. The hierarchal importance of the generated chemical features was investigated by negative feature selection, revealing properties related to cyclization and atom hybridization status to be the most important characteristics for accurate prediction. Our study provides a ready-to-use GT1:acceptor predictor which in addition can be trained on other datasets enabled by the automated feature generation pipelines.

Introduction

Glycosylation is a crucial step to obtain a plethora of biologically and industrially relevant molecules, from proteins to natural products and artificial compounds.[1] Accordingly, glycosylation is one of the most common reactions in the biosphere. However, to achieve the required control of stereo- and regioselectivity, organic chemists apply a succession of reactions, including protecting group manipulations and bond activations, amounting to low chemical yields, poor atom economy, and large amounts of waste.[2,3] In Nature, these reactions are mainly catalysed by glycosyltransferases, enzymes which offer perfect stereoselectivity and often high regioselectivity in a single reaction with unprotected substrates.[4,5] However, the factors governing acceptor specificity and regioselectivity of glycosyltransferase reactions are poorly understood, making it challenging to select an appropriate biocatalyst without extensive experimentation.[6]

Glycosyltransferases are phylogenetically organized into 115 families (as of May 15th, 2023) in the Carbohydrate Active Enzymes (CAZy) database (http://www.cazy.org/).[7] Glycosylation of natural products and secondary metabolites is primarily catalysed by glycosyltransferase family 1 (GT1) enzymes, which thus represent important biocatalysts for biotechnological applications.[1] GT1 enzymes have a GT-B fold, catalysing glycosylation in a cleft between two Rossmann-like domains, the N-terminal domain binding mainly the acceptor substrate(s), and the C-terminal domain binding mainly the α-glycosyl donor.[8] Usually, this glycosyl donor is a uridine diphosphate-activated sugar, and thus GT1s are called UDP-dependent glycosyltransferases or UGTs.[9] They catalyse *C*-, *O*-, *N*- and *S*- glycosylation with an inversion of stereochemistry, leading to β-linked products.[10,11] The reaction proceeds through an oxocarbenium glycosyl intermediate, with the catalytic dyad sharing the abstracted proton.[12] However, while much is known about their structures and mechanisms – 59 GT1 enzymes have at least one deposited crystallographic structure, and 338 are biochemically characterized as of May 15th, 2023 according to the CAZy database – little is known about their acceptor scope, except that it is tremendously varied with thousands of different acceptors being reported, and individual enzymes vary from highly specific to very promiscuous.[13,14] Their activity is difficult to infer from biological data since a single organism can contain over hundred different GT1 genes.[15]

Machine learning (ML) is emerging as a powerful tool in enzymology, due to its strength in recognizing patterns in complex data.[16,17] Accordingly, ML has previously been employed to predict enzyme-substrate specificities.[18] This includes a random forest thiolase activity predictor,[19] a gradient-boosted regression tree capable of predicting the donor specificity of GT-A fold glycosyltransferases,[20] and a random forest adenylate-forming enzyme substrate and function predictor.[21,22] In addition, a decision tree-based algorithm, GT-Predict, has been developed specifically for GT1 enzymes to predict GT1:acceptor pairs.[6] GT-Predict is trained on reactivity measurements of 54 *Arabidopsis thaliana* GT1 enzymes against 91 structurally diverse glycosylation acceptors. GT-Predict was not tested on independent data, and testing on substrates absent from the training set would require the manual addition of substrate features. For sequences outside the training data (*i.e.,* non-*Arabidopsis* GT1 enzymes), GT-Predict returns the substrate reactivity measured experimentally for the closest *A. thaliana* homolog. Given that phylogeny has been shown to be a relatively poor predictor of GT1 specificity,[14] there is potential for further development.

In this study, we aimed to address the broad landscape of GT1:acceptor reactivity by implementing a pan-specific predictor able to process enzymes and chemicals outside the training dataset. We used a random forest architecture trained on 4160 data points (each representing a GT1:acceptor pair) publicly available through the GT-Predict publication.[6] We developed an automated pipeline for enzyme and substrate feature generation, capable of parsing all known GT1 sequences and automatically generating 153 chemical features for any potential acceptor substrate, thereby allowing predictions on all GT1:acceptor pairs (Figure 1). The model, named Glycosyltransferase Acceptor Specificity Predictor (GASP), was tested on an in-house generated independent dataset of 1001 data points, demonstrating the generation of a generic predictor with a balanced accuracy of 72% to evaluate any GT1:acceptor pair. The performance of GASP was compared to baseline models, to GT-predict, to that of a group of GT1 experts for the glycosylation of the plant defensive compound 2,4-dihydroxy-1,4-benzoxazinone (DIBOA), and to random selection for the glycosylation of the essential medicine niclosamide. In addition, negative feature selection was performed to understand the importance of the 153 generated chemical features.

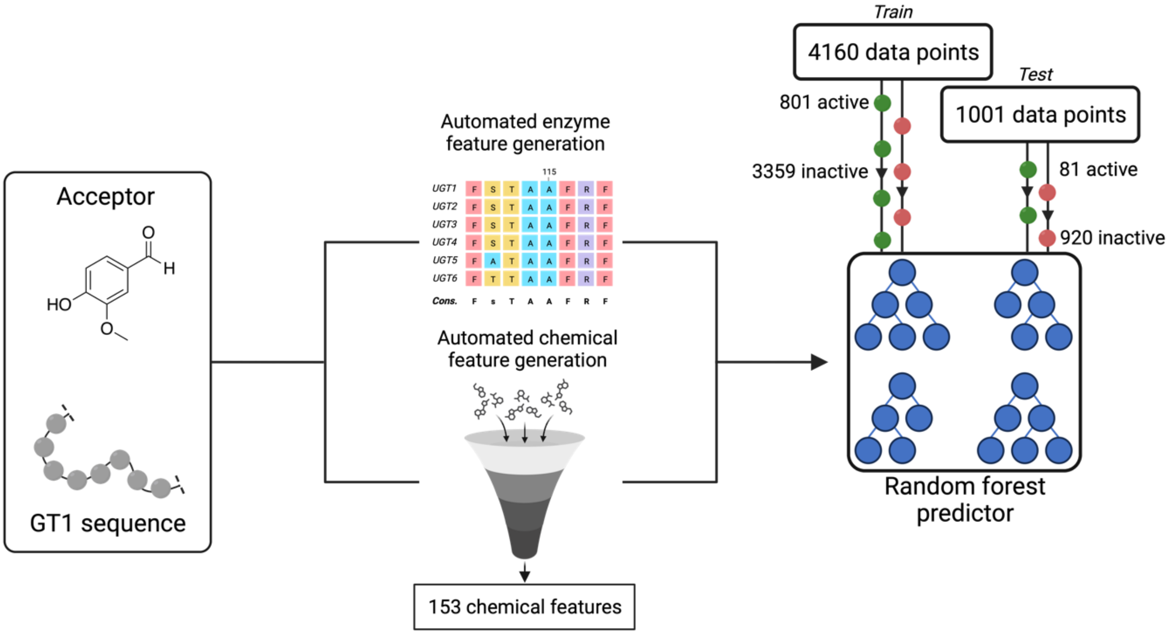


Figure 1. The general concept of GASP: a GT1:acceptor pair consisting of an acceptor and a GT1 sequence is used as input to two automated feature generation pipelines: i) the enzyme feature generation based on an MSA and BLOSUM62 encoding, with colors corresponding to amino acid type, and ii) the substrate feature generation based on chemical features (Figure 2). These features are fed into a random forest predictor, that then returns the predicted reaction probability of the calculated GT1:acceptor pairs. GASP is trained on data from the GT-Predict publication and tested on an independent in-house dataset (active pairs shown as green balls and inactive as red balls).

Methods

Test dataset generation

24 GT1 genes randomly selected from NCBI were synthesized by Genscript (USA) in a modified pET28a(+) vector with an N-terminal 6xHis-tag followed by a TEV-cleavage site and the gene of interest. BL21 Star (DE3) cells (ThermoFisher Scientific, USA) carrying a pET28a(+) vector with the GT1-gene of interest between restriction sites NcoI (5′) and XhoI (3′) were inoculated with 1% (v/v) overnight culture and grown at 37°C until OD_{600} 0.5–0.8 in Luria-Bertani media supplemented with 50 µg/mL kanamycin. Protein expression was induced with 0.5 mM isopropyl-β-D-thiogalactopyranoside, and cells were grown overnight at 18°C. Cells were harvested by centrifugation (4,000 *xg*, 15 min, 4°C) and stored at –20°C. All purification steps were done on ice or in a cold room. Cell pellets were thawed and dissolved in lysis buffer (50 mM HEPES, 300 mM NaCl, 20 mM imidazole, 1 mM dithiothreitol (DTT), pH 7.4, supplemented with 1 µg/mL DNAse I and one cOmplete EDTA-free protease inhibitor cocktail (Roche) tablet per 50 mL lysis buffer). Cells were lysed via three passes through a French press (EmulsiFlex C5, Avestin) and the lysate was cleared by centrifugation (12,000 *x*g, 40 min, 4°C). The supernatant was incubated with Ni-NTA beads (HisPur NiNTA resin, Thermo-Fischer) with gentle shaking (1 h) and the beads were washed three times with wash buffer (50 mM HEPES, 300 mM NaCl, 20 mM Imidazole, pH 7.4). Bound proteins were eluted with elution buffer (50 mM HEPES, 300 mM NaCl, 250 mM Imidazole, pH 7.4). The buffer was exchanged to 50 mM HEPES pH 7.4, 50 mM NaCl, and 2 mM DTT for storage. The protein concentration was adjusted to 5 mg/mL (estimated by A_{280} using a Nanodrop spectrophotometer) when necessary, and aliquots were flash-frozen in liquid nitrogen and stored at –80°C.

Each GT1 enzyme was assayed against a diverse substrate library of compounds representing a typical GT1 acceptor (n=88, Appendix 1) using an in-house developed NADH-coupled enzyme assay in 96-well format; UDP release by the GT1 reaction was detected by coupling it to NADH consumption through the combined action of pyruvate kinase (UDP + phosphoenolpyruvate → pyruvate) and lactate dehydrogenase (pyruvate + NADH → NAD^+ + lactate). The consumption of NADH was followed by A_{340} nm. A 150 µL of reaction mixture consisted of 3 µL of a substrate (10 mM in DMSO), 102 µL of assay buffer (50 mM HEPES, pH 7.4, 50 mM KCl, 5 mM $MgCl_2$, 1 mM EDTA, 1.5 mM DTT, 0.6 mM NADH), 15 µL of detection solution (8 mM phosphoenolpyruvate, 40 U/mL pyruvate kinase, 60 U/mL lactate dehydrogenase), and 15 µL of enzyme. The reaction was initiated by the addition of 15 µL of 10 mM UDP-α-D-glucose (UDP-Glc) and shaken linearly for 10 seconds before reading out A_{340} for 1 hour at 15-second intervals, 25°C, in a Synergy H1 plate reader. Data were analysed with R (https://www.R-project.org/) using RStudio (https://www.RStudio.com). Slopes were fitted (A_{340}/sec), and initial apparent rates were calculated (k_{obs} = slope/[NADH]/[enzyme]). Background activity from enzyme preparations (no substrate added) was subtracted.

Reactivity classification pipeline

A pipeline was constructed for the conversion of reaction rates to reactivity Booleans (*i.e.,* reactive and non-reactive). Reactive GT1:acceptor pairs are identified with outlier detection, since most measurements are of non-reactivity, typically with a sharp contrast to a minor set of non-zero rates (Figure S1). The outlier detection is performed independently on each enzyme by assuming the measurements follow a normal distribution N(μ=0, σ=σ(measurements)), *i.e.,* they are all non-reactive with non-zero rates occurring due to noise. From the distribution, a *p*-value is calculated to quantify how extreme any of the measurements are. Adjusted *p*-values were calculated from the *p*-values with the Holm method. Measurements that have both p-value > 0.05 and adjusted p-value > 0.05 are considered to fit the null hypothesis and are therefore classified as non-reactive observations, while measurements with both p-value < 0.05 and adjusted p-value < 0.05 does not fit the null-hypothesis, so are classified as observations of reactivity. Some data points have a p-value < 0.05 but adjusted p-value > 0.05 which was considered inconclusive evidence; thus those data points were discarded.

Enzyme feature generation pipeline

A pipeline was developed for generating enzyme features that incorporate GT1 enzyme sequences from experimental datasets (*i.e.,* the test dataset, GT-Predict dataset, and reactions from literature) and the CAZy database (26,335 unique Genbank ID entries as of Dec. 2nd, 2021). Sequences from experimental datasets were aligned with MUSCLE[23] and combined with GT1 sequences from CAZy, filtered in length to range from 300 to 600 amino acids. Subsequently, a Hidden Markov Model was built upon the combined set of GT1 sequences using HMMER. Non-consensus positions were discarded, where a consensus position was identified as the majority of sequences containing the same letter for that location. Sequence alignments with less than 80% identity to the consensus sequence (*i.e.,* the sequence with the most frequent amino acids at each position) were discarded, yielding a set of 10,374 sequences. As the N-terminus region is most important for acceptor preference, each of the remaining 10,374 sequences was split in half, and only the part corresponding to the N-terminus was kept for amino acid encoding with BLOSUM62.

Substrate feature generation pipeline

To enable easy prediction of an active GT1 enzyme for any acceptor substrate, we developed a pipeline for substrate feature generation: acceptors represented as PubChem CIDs are converted to SMILES and used as input to RDKit (https://www.rdkit.org) , webchem[24] and E3FP[25] to generate molecular features (Figure 2). Molecular properties are found with the RDKit software and curated from PubChem with the webchem R package.[24] In addition, RDKit is used for generating 3D representations of the chemical compounds in PDB format, which are further used to generate area and volume features with the PyMOL Molecular Graphics System (Version 2.0 Schrödinger, LLC), and ProteinVolume,[26] respectively. E3FP[25] is used for generating molecular fingerprints. The fingerprints are projected into a metric space by applying MultiDimensional Scaling (MDS) to pairwise Euclidean distances calculated between all the molecular fingerprints. Thus, the chemical features from the molecular fingerprints are represented in a 12-dimensional space. MDS was employed to reduce the dimensions of the molecular fingerprints, thereby mitigating the risk of a potential dimensionality problem. A reduction to 12 dimensions was chosen to balance the need for retaining enough information to distinguish different substrates while avoiding fingerprints dominating the substrate encoding. Furthermore, since random forest is employed, it is anticipated that any extraneous MDS features will simply be excluded from the decision trees. Ultimately, all these substrate features are concatenated to a single feature vector.

Model training and evaluation

GT1:acceptor pairs from the GT-Predict dataset (77 chemicals and 73 GT1 enzymes, 4160 datapoints) was encoded using the BLOSUM62 encodings and substrate features as described previously, concatenating them both into a singular feature vector. After removing redundant features with identical values across the entire dataset, the encoded GT-Predict dataset was used to train and optimize a random forest predictor as follows. The effects of "n_estimators" and "max_depth" hyperparameters were first examined manually, and then a more thorough grid search of a larger set of hyperparameters was implemented based on the five-fold cross-validation and area under the receiver operating characteristic curve AUROC (Table S1). Since an exhaustive grid search might lead to overfitting, we decided to keep both the model after manual search and the best performing model after the grid search for the final evaluation on the independent test set.

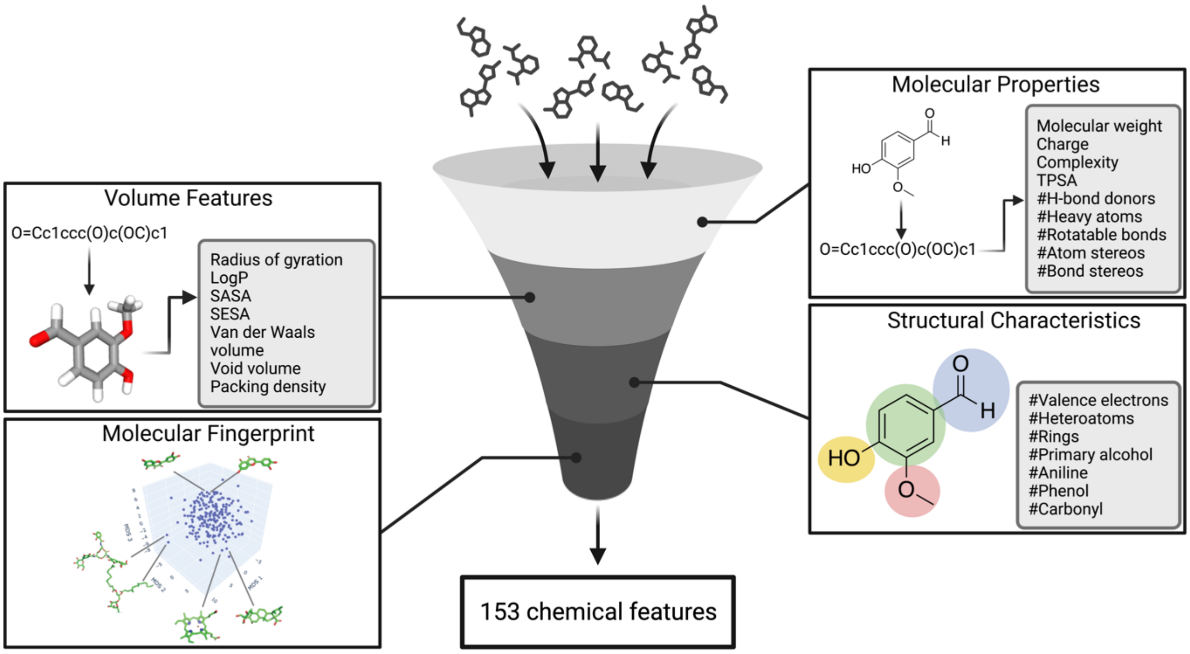


Figure 2. The chemical feature generation pipeline can take CIDs or SMILES and generate chemical features. If a CID is used, SMILES are generated from the CID. Molecular properties are then retrieved from PubChem via webchem[24] using the SMILES. The SMILES are passed to RDKit which creates a molecular representation, including 3D conformers that are written to PDBs and translated to volume features. RDKit is then used to generate structural characteristics, while E3FP[25] is used to generate molecular fingerprints from the SMILES representation (the symbol '#' indicates 'number of'). All pairwise Euclidean distances are calculated between the molecular fingerprints using E3FP which are converted to projected points in a k-dimensional space (here, k=12) using MultiDimensional Scaling (MDS). Features from all steps above are concatenated into a total of 153 chemical features.

The two developed models were tested using an independent in-house dataset (1001 datapoints, see *Test dataset generation*), using the same protocol for feature generation. The AUROC, calculated with the scikit-learn metrics package[27] in Python (version 3.8.5), indicated an overfitting for the best model from the grid search (Figure S2), and consequently, the corresponding model was discarded. The resulting model was further evaluated by the balanced accuracy, precision, recall, and F1-score. The balanced accuracy (eq. 1), precision (eq. 2), recall (eq. 3), and F1-score (eq. 4) were calculated as follows (false negative (FN), false positive (FP), true negative (TN), true positive (TP)):

$$\text{Balanced accuracy} = \frac{\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}}{2}$$

(1)

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

(2)

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

(3)

$$F_1 = 2\,\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

(4)

To calculate the confusion matrix for the reporting purposes, the threshold of 0.345 corresponding to the maximum F1-score was selected (Figure S3). However, the raw score returned by GASP was eventually used in ranking the sequences in the subsequent experimental validation (see *Case study: glycosylation of GASP-predicted GT1s vs expert selection and random selection*).

Comparison to baselines and single task models

To examine the performance of GASP, we constructed baseline and single task models as described by Goldman et al.[18] (Table S2). Specifically, we trained a Levenshtein KNN model, a Tanimoto KNN model, and a Ridge Regression model trained on random features; henceforth denoted the “baseline models”. Due to the limited overlap between the GT-Predict dataset and the in-house data, only eight individual enzyme discovery models and six individual substrate discovery models were constructed (Table S3). In addition to these baseline models, two single task GASP models were constructed – one for enzyme discovery and one for substrate discovery – using the same overlapping enzymes and substrates as the baseline models, denoted as the “single task models”. Finally, the full GASP model was tested on the same subset of GT1:acceptor pairs used to evaluate the enzyme and substrate discovery models. As the full GASP incorporates information about both enzyme and substrate, it is in theory able to learn the interactions between the two, known as a compound-protein interaction (CPI) model.

Comparison to GT-Predict

As a comparison to the performance of GT-Predict model, the leave-one-out validation protocol from the original publication was replicated using our GASP model and the

Arabidopsis thaliana data from GT-Predict (Table S4). The performance was evaluated using accuracy (eq. 5) and Matthews Correlation Coefficient (MCC) (eq. 6):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \quad (5)$$

$$\text{MCC} = \frac{(\text{TP} \cdot \text{TN}) - (\text{FP} \cdot \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}} \quad (6)$$

All hyperparameters of the GASP leave-one-out models were the same as for the full model, as was the threshold chosen for metric calculation. It was impossible to calculate the MCC for 16 substrates due to lack of positive labels in the corresponding subset. The average MCC metric was therefore pruned of these substrates.

Case study: glycosylation of GASP-predicted GT1s vs expert selection and random selection

To test the performance of GASP, a small comparative case study for the glycosylation of DIBOA and niclosamide via expert-selected and GASP-predicted GT1s was carried out. Only GT1s available from our in-house library were considered. For the DIBOA case, expert-selected GT1s were inferred by employing intuition to assess the structural similarity between DIBOA and polyphenols from a publicly available dataset[28] and then choosing among 40 GT1s enzymes that are known to be active on the most similar polyphenol structures, namely 5,7-dihydroxychromone, 4,7-dihydroxycoumarin, 4-methylesculetin, and 4-methyllimetol. GT1s which were active with 3 out of the 4 similar polyphenols were chosen, resulting in six protein sequences. For the selection of GASP-predicted sequences, six GT1s among the highest probability scores present in our stocks were chosen, resulting in a total number of six expert-selected versus six GASP-predicted enzymes (Table S5). GT1 enzymes BX8 (AAL57037.1) and BX9 (AAL57038.1) from *Zea mays* were chosen as positive controls.[29]

Our previous efforts for glycosylation of niclosamide had revealed that 10 out of 19 randomly selected GT1s screened were active, albeit yielding very low amounts of the niclosamide-Glc. For the case study of niclosamide, we therefore examined the performance of GASP to predict GT1s for niclosamide glycosylation. Using the top GASP predictions to construct an initial list of 14 sequences, 2 enzymes with SoluProt[30] scores lower than 0.450 were removed, resulting in a total number of 12 GASP-predicted GT1s.

Selected GT1s were expressed as described in the test dataset generation. Proteins were extracted from 0.5–1 L cell cultures. The filtered supernatant was purified by nickel affinity chromatography (HisTrapTM FF, GE Healthcare, Sweden) on an ÄKTA pure (GE Healthcare, Sweden) system. After concentration and buffer exchange, each GT1 enzyme was assayed for glycosylation activity against DIBOA or niclosamide using UDP-Glc as the donor substrate.

The DIBOA glycosylation reactions were initiated via the addition of 100 µg/mL enzyme to the reaction mixture of 0.5 mM DIBOA from a 50 mM stock in 100% DMSO, 2 mM UDP-Glc in water and 100 mM citrate-phosphate buffer (pH 7.0) in a total reaction volume of 180 µL and incubated for 1 h at 30°C while shaking linearly at 300 rpm. Thirty microliters of the reaction mixture were withdrawn and mixed with 30 µL of methanol to stop the reaction and centrifuged for 10 min to remove any precipitated proteins. Forty microliters of the resulting supernatant were then diluted to 200 µL with MilliQ water before injection into an Ultimate 3,000 Series apparatus equipped with an Agilent ZORBAX Eclipse Plus C18 column. A gradient of solutions A (0.1% aqueous formic acid) and B (100% Acetonitrile) was used as mobile phase for analyte separation at a flow rate of 1 mL/min: gradient increase from 2% B to 70% B between 0–4 min, then immediate increase to 100% B until 4.5 min; and drop to 2% B after 4.5 min until the separation is finished at 5 min. The system was kept at 30°C and DIBOA and DIBOA glycoside were monitored via a UV detector at 220 and 240 nm. Monitoring and data handling were operated using Chromeleon software (Thermofisher).

Glycosylation of niclosamide via GASP-predicted GT1s was carried out in reactions containing 50 µg/mL of each enzyme, 5 mM of UDP-Glc, and < 1 mM niclosamide from a < 7 mM stock in 100% DMSO. Final niclosamide concentrations in the reactions are rough estimations since a significant amount of it could not be solubilized fully in DMSO even at 7 mM. Reactions with a total volume of 100 µL were run in a 50 mM potassium phosphate buffer (pH 7.45) with 50 mM NaCl at 30°C and 300 rpm for 2 h. A hundred microliters of 100% methanol were added to terminate the reactions at the end of 2 h, followed by centrifugation at 2,451 x g for 30 min at 4°C to remove precipitations. Prior to HPLC analysis, 150 µl from the upper phase of each sample were added an equal volume of methanol to facilitate niclosamide solubility further. The HPLC analysis was carried out as described for the DIBOA samples, except for a run time of 9 min and absorbance recording at 290 nm.

For niclosamide glycosylation via randomly selected GT1s, enzymes at varying concentrations were reacted with an undetermined amount of niclosamide and 3 mM UDP-Glc in a buffer containing 50 mM HEPES and 50 mM NaCl (pH 7.0) overnight at 30°C.

Chemical feature selection

To compare the importance of the 153 generated chemical substrate features, feature selection was performed. Individual features were deselected iteratively, where predictive performance was measured after temporarily leaving out each remaining feature. The feature whose removal led to the smallest decrease in performance was then left out permanently for further iterations until only one remained, which may be considered the most important single feature in discerning reactivity from non-reactivity. At each iteration, the available data points were randomly split into train and test sets, where the test set contained 20% of substrates. These were selected by randomly picking a single substrate, and then finding its nearest neighbors based on the highest correlation on their chemical feature values. Performance metrics were averaged between 10 repetitions of each iteration.

The performance for each deselection was evaluated by a custom metric, named topP, which is designed to minimize false positives. This is motivated by predictor application,

where experiments will only be carried out on the top-scoring predictions. Thus, this metric has a bias for the accuracy of top-scoring candidates rather than equal weight for all GT1:substrate pairs. TopP is defined by assigning weights from 1 to P to the top P predictions in ascending order, where P is the number of positives (reactive pairs). TopP is then equal to the sum of weights given to true positives, after normalization.
Moreover, as the MDS features are abstract values not representing a single chemical property, their use requires additional justification. Consequently, we studied their importance by training GASP models without any of the 12 MDS features and comparing the resulting performance to the full GASP model.

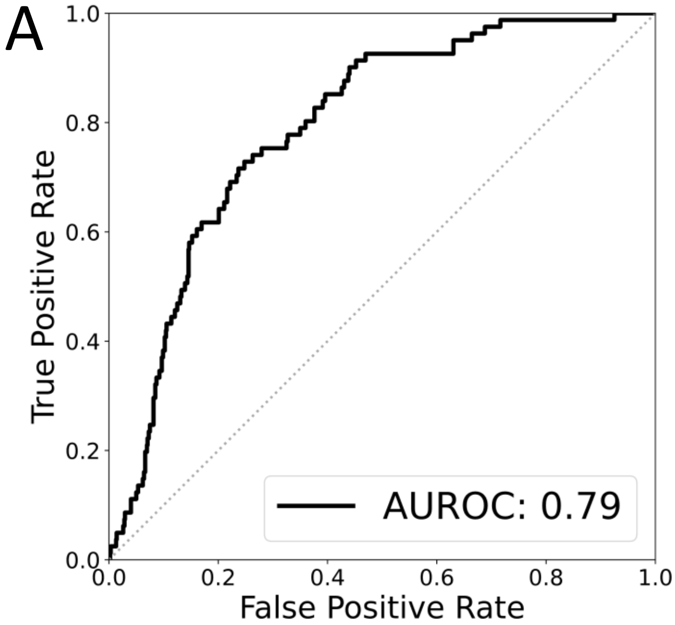
Results

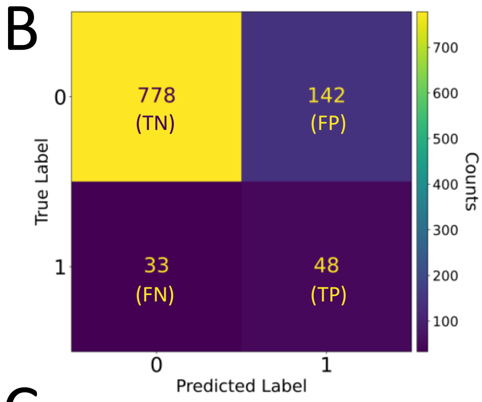
Test dataset

For independent validation of predictor performance, a test dataset was collected by measuring initial rates (k_{app}) of 24 GT1 enzymes from 15 different plants on 88 acceptors. This yielded a total of 1031 data points (not all acceptors were tested against all enzymes) of which 81 were active, 920 were inactive, and 30 were inconclusive. The inconclusive data points were removed from the dataset yielding a total of 1001 data points with a distribution of 8% active and 92% inactive GT1:acceptor pairs (see "dataset1.xlsx" in supplemental data).

Algorithm generation and evaluation

The outputs of our enzyme and substrate feature generation pipelines are fed to a random forest classifier consisting of 1,000 trees. We refer to this as the GASP model. It was trained on a curated published dataset of 4,160 data points, which were reactivity measurements between 77 chemicals and 73 GT1 enzymes (53 from *Arabidopsis thaliana*, 10 from *Lycium barbarum*, 6 from *Avena strigosa*, 2 from *Medicago truncatula*, 1 from *Streptomyces antibioticus*, and 1 from *Vitis vinifera*).[6] GASP was subsequently tested on the independent in-house test dataset, with the predicted probabilities covering the full range of values (Figure S4). Here, the random forest predictor achieved an AUROC of 0.79 (where an AUROC of 0.5 indicates random guessing and a value of 1.0 indicates perfect classification) (Figure 3A). Interestingly, the performance does not appear to be determined solely by similarity to the training data, as observed when examining the performance from enzymes belonging to the same organisms (Figure S5). With a probability threshold of 0.345 corresponding to the maximum F_1-score of 0.30, a confusion matrix was calculated (Figure 3B), with a precision and recall of 0.25 and 0.59, respectively (Figure 3C). We observed a high number of false positives compared to true positives, probably due to the imbalance of labels in the test data, as the majority of the GT1:acceptor pairs are inactive (Figure 1). If the confusion matrix is normalized by the number of points in each class, we instead observe that only 15% of the inactive GT1:acceptor pairs are falsely predicted as reactive, while 85% are predicted correctly (Figure S6). A balanced accuracy of 72% was obtained, although it should be noted that by lowering the threshold to 0.265, GASP can obtain the maximum balanced accuracy of 74% (Figure S7).





C

Balanced Accuracy	Precision	Recall	F_1
72%	0.25	0.59	0.30

Figure *3.* **A** ROC curve for GASP predictions on the in-house dataset (black line) with the corresponding AUROC value. The grey dotted line corresponds to the random predictor. **B** Confusion matrix and **C** calculated test metrics of the GASP model on test dataset using the probability threshold of 0.345 maximizing the F1 score.

Comparison of GASP and alternative models

First, we validated the GASP architecture by following the protocol described by Goldman et al.[18], constructing baseline and single task models for both enzyme discovery and substrate discovery for the enzyme and substrate subsets with sufficient data (see *Comparison to baselines and single task models* in Methods). We observed a significant increase in performance between the full GASP model and all baseline models (Figure S8). Interestingly, the full model exhibited similar performance to the single task GASP models within one standard deviation, indicating that the CPI nature of the full GASP model does not produce higher performance in the setting when sufficient experimental data for a given substrate or enzyme are available. This aligns with the conclusions by Goldman et al.[18] However, incorporating both enzyme and substrate features into the model did not compromise its performance and also enabled the full GASP model to predict new GT1:acceptor pairs without the need to collect sufficient training data and retrain a new single task model.

We also compared GASP to the previously published GT-Predict model.[6] Due to the nature of the GT-Predict architecture, we were unable to use our in-house dataset to test GT-Predict. Instead, we replicated their leave-one-out validation (see *Comparison to GT-Predict* in Methods). For both the average accuracy and average MCC score, the two models lie within one standard deviation of each other, and a two-sided t-test reveal them to be statistically similar (*p*-value of 0.918 and 0.227 for the accuracies and MCC scores, respectively). This indicates that in the GT-predict setting, the models have equal performance. However, the pan-specificity unique to GASP allows it to automatically generate features and make predictions for new GT1:acceptor pairs, which is a major practical benefit.

DIBOA glycosylation by expert-selected versus predicted GT1s

DIBOA is one of the most common benzoxazinoids in plants, taking part in plant defence. It is stored in the vacuole in its glycosylated form to reduce autotoxicity. Upon cell damage, a β-glucosidase hydrolyses the glycoside to release the toxic aglycon in response to pest or pathogen attack.[31] DIBOA is of interest as a phytoremediation agent due to its ability to degrade the recalcitrant herbicide atrazine,[32] and as a biopesticide due to its toxicity to pests and pathogens. There is only limited knowledge of GT1 enzymes active on DIBOA, and thus it is interesting to discover novel DIBOA-glycosylating enzymes.

BX8 and BX9 are two well-characterized GT1s that are known to glycosylate DIBOA,[29] thus were chosen as positive controls in this study. The DIBOA molecule carries two potential glycosylation sites, and our results indicate that while BX8 and BX9 produce each a single product, they present different regioselectivities as seen in two separate peaks with different retention times on HPLC spectra (Figure S10).

To discover novel DIBOA-glycosylating enzymes, we leveraged an in-house dataset of 40 GT1s reactivity on different polyphenols.[28] Based on DIBOA's chemical similarity to some of the substrates in this dataset (5,7-dihydroxychromone, 4,7-dihydroxycoumarin, 4-methylesculetin, and 4-methyllimetol), we selected six in-house GT1 enzymes to be assayed for DIBOA activity (referred to as "expert selection"). In parallel, we predicted DIBOA-active GT1 enzymes using GASP (Figure S11) and chose six of the top-ranking enzymes present in our stock (see *Case study: glycosylation of GASP-predicted GT1s vs expert selection and random selection* in Methods). As summarized in Table S5, five out of six expert-selected GT1s showed activity on DIBOA, while for the GASP-predicted GT1s, the success rate was three out of six. Among expert-selected GT1s, only *Rh*Gt1 from *Rosa hybrid* was inactive. As for the remaining five, only GT171E5 from *Carthamus tinctorius* produced the same product as the BX9 enzyme, while the others showed the same product as BX8 (Figure S12). As the in-house dataset does not provide any information about the regioselectivity of the reactive GT1:acceptor pairs, GASP is unable to predict this property. Nevertheless, a similar trend to the expert-selected GT1s was observed for the three active algorithm-predicted GT1s, namely GT184A57 from *Eutrema japonicum*, GT174F2 from *Arabidopsis thaliana*, and GT175L5 from *Lycium barbarum*, which all produced the same product as BX8 (Figure S13). It should be noted that the commercial DIBOA preparation used as a standard contained trace amounts of a compound with the same retention time as that produced by BX8, as can be seen in the HPLC spectra of the negative control samples. The corresponding peak area was subtracted.

Niclosamide glycosylation by random in-house versus predicted GT1s

Niclosamide is a lipophilic and weakly acidic salicylanilide widely used as an anti-helminth drug for the treatment of tapeworm infections.[33] Unfortunately, niclosamide's poor aqueous solubility reduces its bioavailability, which presents a major challenge for the realization of its pharmaceutical potential.[34] Glycosylation can be a powerful tool to increase the aqueous solubility of such compounds. Our previous random screening of in-house GT1 enzymes for niclosamide glycosylation had identified 10/19 (53%) active enzymes (Table S6), although the activities were very low, and conversion yields were too low to quantify.

Hence, we employed GASP to predict efficient niclosamide-glycosylating GT1s (Figure S14). From the 12 sequences assessed, five could not be expressed in *E. coli*, and one was expressed in its insoluble form (Table S6). Five out of six remaining sequences, however, demonstrated significant niclosamide glycosylation activity as seen in the HPLC spectra (Figure S15). The GASP hit rate for the niclosamide case was thus 83% (5 out of 6).

Acceptor features important for prediction performance

To learn which of the 153 chemical features describing the acceptors were more important to prediction performance, we performed negative feature selection. The ten most important chemical features from the negative feature selection are shown in Table 1, where chemical features relating to atom hybridization and cyclic properties (*i.e.,* number of saturated rings, aromatic rings, furan structures and aromatic nitrogens) are predominant. Indeed, the fraction of sp^3 hybridized carbons in a molecule is the most important feature, while also impacting the features ranked 4th, 6th, and 10th. The hybridization of nitrogen impacts features 7th and 8th. Since GT1s predominantly glycosylate polyphenolic compounds, and GASP was trained primarily on these compounds, it is compelling to observe that the performance depends on the description of cyclic structures.

It is worth noting that the negative feature selection ranks the chemical features based on their importance to achieve high accuracy, not whether these features favor glycosylation. Indeed, while the number of sulfide bonds (*i.e.,* thioether) was ranked as the fifth most important feature, these were only present in three out of the 88 chemicals with none of them showing reactivity in 82 reactions.

To evaluate the usefulness of the MDS fingerprint reduction included in the chemical features, we evaluated the model's performance without its use: when removing all MDS values from the substrate feature set, we observed a decrease in prediction performance (Figure S16). Together with a dimension of the MDS-generated space being the second most important feature, we conclude that the molecular fingerprints serve as relevant features for improving the model's performance, and the dimensionality reduction conserves useful information.

Table 1. The ten most important features found from the negative feature selection (NPR: normalized principal moment ratio, MDS: multidimensional scaling).

Order	**1**	**2**	**3**	**4**	**5**
Chemical feature	Fraction sp^3 carbons	MDS 9	No. of valence electrons	No. of saturated rings	No. of sulfide bonds

Order	**6**	**7**	**8**	**9**	**10**
Chemical feature	No. of furans	No. of Quaternary nitrogens	No. of aromatic nitrogens	NPR1	No. of aromatic rings

Discussion

In this work, we demonstrated the synergistic effect of high-throughput data generation with a chemically informed machine learning predictor. Indeed, we proposed GASP, an enzyme specificity predictor trained on the largest experimental dataset on GT1 enzymes which performs well on enzymes and acceptors absent from the training set. This was demonstrated using an independent test dataset of 1001 datapoints, where GASP outperformed all baseline models. A leave-one-out comparison to the previous state-of-the-art model for predicting GT1:acceptor pairs, GT-Predict[6], revealed a statistically similar performance, demonstrating the potential of GASP. And while the full model also exhibited similar performance to single task models, the pan-specificity of GASP allows it to readily incorporate and predict new GT1:acceptor pairs. This is observed when we examined the performance of enzymes from individual organisms, where predictions on proteins from organisms absent from the training data showed good performance even when the phylogenetic similarity with *Arabidopsis thaliana* – which comprises the majority of the training data – was low. The model thereby exhibited the ability to accurately extrapolate beyond the training GT1:acceptor pairs, enabling researchers to estimate the substrate activity of new GT1 enzymes without requiring preliminary experimental analysis. It should be noted that the enzyme feature generation pipeline requires alignment of new sequences to the current consensus sequence, and sequences with very low similarity might result in a drop in performance.

To examine this application of GASP, we conducted two use case studies with DIBOA and niclosamide. GASP significantly outperformed a random selection of GT1s for the niclosamide case, as GASP had a hit rate of 83% compared to the 53% obtained with random selection. In the DIBOA case, a hit rate of 50% for the GASP-selected enzymes indicates that – not surprisingly – GASP cannot compete with highly trained researchers in the field, who got a hit rate of 83%. However, GASP can parse a much larger number of sequences, including never-assayed sequences, while expert selection is limited to sequences evaluated against analogues. In conclusion, these case studies show that GASP can be utilized as a tool for preliminary assessment of enzymes.

It is particularly interesting that GASP is successful despite the fact that enzyme features are generated with multiple sequence alignment, and therefore the algorithm does not directly use such important characteristics as loops of varying length near the active site, which are known to have a strong impact in CAZymes' specificity, including GT1s'.[35] With the recent release of AlphaFold2[36] and the wealth of accurate structural models it provides, it might be feasible to incorporate structural information of the overall protein fold as well as active site loops, similar to what has been done for the predictions of binding parameters of cellulases.[37] In addition to incorporating structural information, future models should address the issue of regioselectivity. While GASP only focused on predicting the acceptor specificity – partially due to the lack of the regiochemical outcome of GT1 glycosylation information in both our datasets and most of the literature – regioselectivity is an important property of the GT1 enzymes. ML models able to predict regioselectivity would thus be highly advantageous when selecting an appropriate GT1 for biocatalysis.

Finally, the developed pipelines enable the addition of new data, thus the present framework can be extended for generating new improved models on other data or in combination with the data used in this work. The provided pipelines for automated feature generation on proteins and chemicals can even be used for other enzyme classes. Furthermore, the in-house dataset employed in this study offers a new, cleaned, and independent GT1 activity dataset for use as training or test sets for future ML models.

Data availability

All activity datasets used herein are included in a supplemental zip file, and GASP code is available at https://github.com/degnbol/GASP.

Acknowledgements

We thank The Novo Nordisk Foundation for supporting this work through grants NNF18OC0034744, NNF10CC1016517, and NNF20CC0035580. This work was also supported by Czech Ministry of Education, Youth and Sports [CETOCOEN Excellence CZ.02.1.01/0.0/0.0/17_043/0009632, ESFRI RECETOX RI LM2023069, ESFRI ELIXIR LM2023055] and the CETOCOEN EXCELLENCE Teaming 2 project supported by Horizon2020 of the European Union [857560].

The authors thank Tiia Kittilä and Folmer Fredslund for selecting sequences for the in-house dataset and Valeria Della Gala for preliminary work on DIBOA.

Figures 1 and 2 were created with BioRender.com.

Declaration of competing interest

The authors declare no competing interests.

References

(1) Nidetzky, B.; Gutmann, A.; Zhong, C. Leloir Glycosyltransferases as Biocatalysts for Chemical Production. *ACS Catal* **2018**, *8* (7), 6283–6300.

(2) De Roode, B. M.; Franssen, M. C. R.; Van Der Padt, A.; Boom, R. M. Perspectives for the Industrial Enzymatic Production of Glycosides. *Biotechnol Prog* **2003**, *19* (5), 1391–1402. https://doi.org/10.1021/BP030038Q.

(3) Desmet, T.; Soetaert, W.; Bojarová, P.; Kařen, V.; Dijkhuizen, L.; Eastwick-Field, V.; Schiller, A. Enzymatic Glycosylation of Small Molecules: Challenging Substrates Require Tailored Catalysts. *Chemistry – A European Journal* **2012**, *18* (35), 10786–10801. https://doi.org/10.1002/CHEM.201103069.

(4) Bowles, D.; Isayenkova, J.; Lim, E. K.; Poppenberger, B. Glycosyltransferases: Managers of Small Molecules. *Curr Opin Plant Biol* **2005**, *8* (3), 254–263. https://doi.org/10.1016/J.PBI.2005.03.007.

(5) Lim, E. K.; Ashford, D. A.; Hou, B.; Jackson, R. G.; Bowles, D. J. Arabidopsis Glycosyltransferases as Biocatalysts in Fermentation for Regioselective Synthesis of Diverse Quercetin Glucosides. *Biotechnol Bioeng* **2004**, *87* (5), 623–631. https://doi.org/10.1002/BIT.20154 .

(6) Yang, M.; Fehl, C.; Lees, K. V.; Lim, E. K.; Offen, W. A.; Davies, G. J.; Bowles, D. J.; Davidson, M. G.; Roberts, S. J.; Davis, B. G. Functional and Informatics Analysis Enables Glycosyltransferase Activity Prediction. *Nat Chem Biol* **2018**, *14* (12), 1109–1117.

(7) Drula, E.; Garron, M. L.; Dogan, S.; Lombard, V.; Henrissat, B.; Terrapon, N. The Carbohydrate-Active Enzyme Database: Functions and Literature. *Nucleic Acids Res* **2022**, *50* (D1), D571–D577. https://doi.org/10.1093/NAR/GKAB1045.

(8) Bidart, G. N.; Putkaradze, N.; Fredslund, F.; Kjeldsen, C.; Ruiz, A. G.; Duus, J. Ø.; Teze, D.; Welner, D. H. Family 1 Glycosyltransferase UGT706F8 from *Zea Mays* Selectively Catalyzes the Synthesis of Silibinin 7-O-β-D-Glucoside. *ACS Sustain Chem Eng* **2022**. https://doi.org/10.1021/acssuschemeng.1c07593.

(9) Ross, J.; Li, Y.; Lim, E.-K.; Bowles, D. J. Higher Plant Glycosyltransferases. *Genome Biol* **2001**, *2* (2), reviews3004.1-3004.6.

(10) Tegl, G.; Nidetzky, B. Leloir Glycosyltransferases of Natural Product C-Glycosylation: Structure, Mechanism and Specificity. *Biochem Soc Trans* **2020**, *48* (4), 1583–1598. https://doi.org/10.1042/BST20191140.

(11) Lairson, L. L.; Henrissat, B.; Davies, G. J.; Withers, S. G. Glycosyltransferases: Structures, Functions, and Mechanisms. **2008**, *77*, 521–555. https://doi.org/10.1146/ANNUREV.BIOCHEM.76.061005.092322.

(12) Teze, D.; Coines, J.; Fredslund, F.; Dubey, K. D.; Bidart, G. N.; Adams, P. D.; Dueber, J. E.; Svensson, B.; Rovira, C.; Welner, D. H. *O*-/*N*-/*S*-Specificity in Glycosyltransferase Catalysis: From Mechanistic Understanding to Engineering. *ACS Catal* **2021**, *11* (11), 1810–1815.

(13) He, J.; Zhao, P.; Hu, Z.; Liu, S.; Kuang, Y.; Zhang, M.; Li, B.; Yun, H.; Qiao, X. Molecular and Structural Characterization of a Promiscuous *C*-Glycosyltransferase from *Trollius Chinensis*. *Angew. Chem. Int. Ed.* **2019**, *131* (131), 11637–11644. https://doi.org/10.1002/ange.201905505.

(14) Zhang, L.; Wang, D.; Zhang, P.; Wu, C.; Li, Y. Promiscuity Characteristics of Versatile Plant Glycosyltransferases for Natural Product Glycodiversification.

ACS Synth Biol **2022**, *11* (11), 812–819. https://doi.org/10.1021/acssynbio.1c00489.

(15) Ross, J.; Li, Y.; Lim, E.-K.; Bowles, D. J. Higher Plant Glycosyltransferases. *Genome Biology 2001 2:2* **2001**, *2* (2), 1–6. https://doi.org/10.1186/GB-2001-2-2-REVIEWS3004.

(16) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-Learning-Guided Directed Evolution for Protein Engineering. *Nat Methods* **2019**, *16* (August), 687–694. https://doi.org/10.1038/s41592-019-0496-6.

(17) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal* **2020**, *10* (10), 1210–1223. https://doi.org/10.1021/acscatal.9b04321.

(18) Goldman, S.; Das, R.; Yang, K. K.; Coley, C. W. Machine Learning Modeling of Family Wide Enzyme-Substrate Specificity Screens. *PloS Comput Biol* **2022**, *18* (2), e1009853. https://doi.org/10.1371/JOURNAL.PCBI.1009853.

(19) Robinson, S. L.; Smith, M. D.; Richman, J. E.; Aukema, K. G.; Wackett, L. P. Machine Learning-Based Prediction of Activity and Substrate Specificity for OleA Enzymes in the Thiolase Superfamily. *Synth Biol* **2020**, *5* (1). https://doi.org/10.1093/SYNBIO/YSAA004.

(20) Taujale, R.; Venkat, A.; Huang, L. C.; Zhou, Z.; Yeung, W.; Rasheed, K. M.; Li, S.; Edison, A. S.; Moremen, K. W.; Kannan, N. Deep Evolutionary Analysis Reveals the Design Principles of Fold a Glycosyltransferases. *Elife* **2020**, *9*. https://doi.org/10.7554/ELIFE.54532.

(21) Robinson, S. L.; Terlouw, B. R.; Smith, M. D.; Pidot, S. J.; Stinear, T. P.; Medema, M. H.; Wackett, L. P. Global Analysis of Adenylate-Forming Enzymes Reveals β-Lactone Biosynthesis Pathway in Pathogenic Nocardia. *Journal of Biological Chemistry* **2020**, *295* (44), 14826–14839. https://doi.org/10.1074/JBC.RA120.013528.

(22) Feehan, R.; Montezano, D.; Slusky, J. S. G. Machine Learning for Enzyme Engineering, Selection and Design. *Protein Engineering, Design and Selection* **2021**, *34*, 1–10. https://doi.org/10.1093/PROTEIN/GZAB019.

(23) Edgar, R. C. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res* **2004**, *32* (5), 1792–1797. https://doi.org/10.1093/NAR/GKH340.

(24) Szöcs, E.; Stirling, T.; Scott, E. R.; Scharmüller, A.; Schäfer, R. B. Webchem: An R Package to Retrieve Chemical Information from the Web. *J Stat Softw* **2020**, *93*, 1–17. https://doi.org/10.18637/JSS.V093.I13.

(25) Axen, S. D.; Huang, X. P.; Cáceres, E. L.; Gendelev, L.; Roth, B. L.; Keiser, M. J. A Simple Representation of Three-Dimensional Molecular Structure. *J Med Chem* **2017**, *60* (17), 7393–7409. https://doi.org/10.1021/ACS.JMEDCHEM.7B00696.

(26) Chen, C. R.; Makhatadze, G. I. ProteinVolume: Calculating Molecular van Der Waals and Void Volumes in Proteins. *BMC Bioinformatics* **2015**, *16* (1), 1–6. https://doi.org/10.1186/S12859-015-0531-2.

(27) Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Thirion, B.; Grisel, O.; Dubourg, V.; Passos, A.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12* (85), 2825–2830.

(28) Prospective author list: Ruben M. de Boer; Dovydas Vaitkus; Kasper Enemark-Rasmussen; Sören Maschmann; Lluís Raich; David Teze; Ditte H. Welner. Prospective Title: Regioselective Glycosylation of Dihydroxycoumarin Derivatives by Family 1 Glycosyltransferases: Experiments and Simulations. **2023**.

(29) Von Rad, U.; Hüttl, R.; Lottspeich, F.; Gierl, A.; Frey, M. Two Glucosyltransferases Are Involved in Detoxification of Benzoxazinoids in Maize. *The Plant Journal* **2001**, *28* (6), 633–642. https://doi.org/10.1046/J.1365-313X.2001.01161.X.

(30) Hon, J.; Marusiak, M.; Martinek, T.; Kunka, A.; Zendulka, J.; Bednar, D.; Damborsky, J. SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics* **2021**, *37* (1), 23–28. https://doi.org/10.1093/bioinformatics/btaa1102.

(31) Frey, M.; Schullehner, K.; Dick, R.; Fiesselmann, A.; Gierl, A. Benzoxazinoid Biosynthesis, a Model for Evolution of Secondary Metabolic Pathways in Plants. *Phytochemistry* **2009**, *70* (15–16), 1645–1651. https://doi.org/10.1016/J.PHYTOCHEM.2009.05.012.

(32) Willett, C. D.; Lerch, R. N.; Lin, C. H.; Goyne, K. W.; Leigh, N. D.; Roberts, C. A. Benzoxazinone-Mediated Triazine Degradation: A Proposed Reaction Mechanism. *J Agric Food Chem* **2016**, *64* (24), 4858–4865. https://doi.org/10.1021/ACS.JAFC.6B01017.

(33) Pearson, R. D.; Hewlett, E. L. Niclosamide Therapy for Tapeworm Infections. *Ann Intern Med* **1985**, *102* (4), 550. https://doi.org/10.7326/0003-4819-102-4-550.

(34) Needham, D. The PH Dependence of Niclosamide Solubility, Dissolution, and Morphology: Motivation for Potentially Universal Mucin-Penetrating Nasal and Throat Sprays for COVID19, Its Variants and Other Viral Infections. *Pharm Res* **2022**, *39* (1), 115–141. https://doi.org/10.1007/s11095-021-03112-x.

(35) Brazier-Hicks, M.; Offen, W. A.; Gershater, M. C.; Revett, T. J.; Lim, E. K.; Bowles, D. J.; Davies, G. J.; Edwards, R. Characterization and Engineering of the Bifunctional N- and O-Glucosyltransferase Involved in Xenobiotic Metabolism in Plants. *Proceedings of the National Academy of Sciences* **2007**, *104* (51), 20238–20243.

(36) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(37) Schaller, K. S.; Kari, J.; Borch, K.; Peters, H. J.; Westh, P. Binding Prediction of Multi-Domain Cellulases with a Dual-CNN. *arXiv : 2207 . 02698v1 [physics . bio-ph] 6 Jul 2022.*